## 作业讲解

作业1，WordCount开发

val lines = List("hello world","hello spark")

val wordlist = lines.flatMap(line => line.split(" ")).map(word =>

(word,1))

wordlist.groupBy(\_.\_1).map{

case(word,list)=>(word,list.size)

}.foreach(println)

作业2：

import org.apache.spark.storage.StorageLevel

import org.apache.spark.{SparkConf, SparkContext}

object VisitCount {

def main(args: Array[String]): Unit = {

val conf = new SparkConf()

val sc = new SparkContext(conf)

val linesRDD = sc.textFile("E:\\testdata")

.filter(line => line.length > 0)

.map { line =>

val arr = line.split("\t")

val date = arr(17).substring(0, 10)

val provinceId = arr(23)

val url = arr(1)

val guid = arr(5)

(date + "\_" + provinceId, (guid, url))

}.filter(line => line.\_2.\_2.length > 5) // length(url) > 5

.persist(StorageLevel.MEMORY\_ONLY)

linesRDD.map(line => (line.\_1 + "\_" + line.\_2.\_1, 1)) //

(date\_provinceId\_guid, 1)

.reduceByKey(\_ + \_) // (date\_provinceId\_guid, pv)

.map { i =>

val arr = i.\_1.split("\_")

val dateProvinceId = (arr(0), arr(1))

val guidCount = 1

val pv = i.\_2

(dateProvinceId, (guidCount, pv))

}.reduceByKey((x, y) => (x.\_1 + y.\_1, x.\_2 + y.\_2))

.map { i =>

val arr = i.\_1.split("\_")

val date = if (arr.length >= 1) arr(0) else "Nil"

val provinceId = if (arr.length >= 2) arr(1) else "Nil"

val pv = i.\_2.\_2

val uv = i.\_2.\_1

(date, provinceId, pv, uv)

}.foreach(println)

}

}

罗剑平 20:55:03

python sklearn对参数调优可以进行网格搜索，不知spark mllib里面有无对应的api？

答：不晓得。

罗剑平 20:58:43

如对lambda，delta分别从（10，100，0.1），从10到100，0.1的步长，给一个评估的函数或参数找到一个或者多个最优参数

答：确实是个思路。

吴卓钧 21:10:38

公安大数据聚类出来4个组，怎么知道哪个组对应涉毒指数高，哪个组对应涉毒指数低？

### 关联算法

通过关联算法生成关联规则，保存到数据库中供以后使用。

关联规则如：啤酒——>尿布

啤酒，尿布——>牛奶

牛奶——>啤酒

郑博 21:58:52

壹号店的推荐是用这个算法去做推荐的吗

答：非常多和杂。